SneakyPrompt: Jailbreaking Text-to-image Generative Models

Year: 2023 | Citations: 141 | Authors: Yuchen Yang, Bo Hui, Haolin Yuan, N. Gong, Yinzhi Cao

Abstract

Text-to-image generative models such as Stable Diffusion and DALL•E raise many ethical concerns due to the generation of harmful images such as Not-Safe-for-Work (NSFW) ones. To address these ethical concerns, safety filters are often adopted to prevent the generation of NSFW images. In this work, we propose SneakyPrompt, the first automated attack framework, to jailbreak text-to-image generative models such that they generate NSFW images even if safety filters are adopted. Given a prompt that is blocked by a safety filter, SneakyPrompt repeatedly queries the text-to-image generative model and strategically perturbs tokens in the prompt based on the query results to bypass the safety filter. Specifically, SneakyPrompt utilizes reinforcement learning to guide the perturbation of tokens. Our evaluation shows that SneakyPrompt successfully jailbreaks DALL•E 2 with closed-box safety filters to generate NSFW images. Moreover, we also deploy several state-of-the-art, open-source safety filters on a Stable Diffusion model. Our evaluation shows that SneakyPrompt not only successfully generates NSFW images, but also outperforms existing text adversarial attacks when extended to jailbreak text-to-image generative models, in terms of both the number of queries and qualities of the generated NSFW images. SneakyPrompt is open-source and available at this repository: https://github.com/Yuchen413/text2image_safety.